

SOLR Statistics

DSpace uses the Apache SOLR application underlying the statistics. SOLR enables performant searching and adding to vast amounts of (usage) data. Unlike previous versions, enabling statistics in DSpace does not require additional installation or customization. All the necessary software is included.

- 1 What is exactly being logged ?
 - 1.1 Common stored fields for all usage events
 - 1.2 Unique stored fields for bitstream downloads
 - 1.3 Unique stored fields for search queries
 - 1.4 Unique stored fields for workflow events
- 2 Disabling Tracking of Statistics
- 3 Web User Interface Elements
 - 3.1 Pageview and Download statistics
 - 3.1.1 Home page
 - 3.1.2 Community home page
 - 3.1.3 Collection home page
 - 3.1.4 Item home page
 - 3.2 Search Query Statistics
 - 3.3 Workflow Event Statistics
- 4 Architecture
- 5 Configuration settings for Statistics
 - 5.1 Pre-1.6 Statistics settings
- 6 Statistics Administration
 - 6.1 Converting older DSpace logs into SOLR usage data
 - 6.2 Statistics Client Utility
 - 6.3 Anonymizing Statistics
- 7 Custom Reporting - Querying SOLR Directly
 - 7.1 Resources
 - 7.2 Examples
 - 7.2.1 Top downloaded items by a specific user
- 8 Managing the City Database File

What is exactly being logged ?

After the introduction of the SOLR Statistics logging, every pageview and file download is logged in a dedicated SOLR statistics core.

In addition to the already existing logging of pageviews and downloads, DSpace also logs search queries users enter in the DSpace search dialog and workflow events.

DSpace 7.x and 8.x does not yet support all features

In DSpace 7.x & 8.x, only usage statistics (pageview, downloads) are logged. Search statistics and workflow reports (which were available in 6.x and below) are not yet supported. See their related tickets: <https://github.com/DSpace/DSpace/issues/2880> and <https://github.com/DSpace/DSpace/issues/2851>

Workflow Events logging

Only workflow events, initiated and executed by a physical user are being logged. Automated workflow steps or ingest procedures are currently **not** being logged by the workflow events logger.

The logging happens at the server side, and doesn't require a javascript like Google Analytics does, to provide usage data. Definition of which fields are to be stored happens in the file **dspace/solr/statistics/conf/schema.xml**.

Although they are stored in the same index, the stored fields for views, search queries and workflow events are different. A new field, `statistics_type` determines which kind of a usage event you are dealing with. The three possible values for this field are **view**, **search** and **workflow**.

```
<field name="statistics_type" type="string" indexed="true" stored="true" required="true" />
```

Common stored fields for all usage events

```

<field name="type" type="integer" indexed="true" stored="true" required="true" />
<field name="id" type="integer" indexed="true" stored="true" required="true" />
<field name="ip" type="string" indexed="true" stored="true" required="false" />
<field name="time" type="date" indexed="true" stored="true" required="true" />
<field name="epersonid" type="integer" indexed="true" stored="true" required="false" />
<field name="continent" type="string" indexed="true" stored="true" required="false"/>
<field name="country" type="string" indexed="true" stored="true" required="false"/>
<field name="countryCode" type="string" indexed="true" stored="true" required="false"/>
<field name="city" type="string" indexed="true" stored="true" required="false"/>
<field name="longitude" type="float" indexed="true" stored="true" required="false"/>
<field name="latitude" type="float" indexed="true" stored="true" required="false"/>
<field name="owningComm" type="integer" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="owningColl" type="integer" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="owningItem" type="integer" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="dns" type="string" indexed="true" stored="true" required="false"/>
<field name="userAgent" type="string" indexed="true" stored="true" required="false"/>
<field name="isBot" type="boolean" indexed="true" stored="true" required="false"/>
<field name="referrer" type="string" indexed="true" stored="true" required="false"/>
<field name="uid" type="uuid" indexed="true" stored="true" default="NEW" />
<field name="statistics_type" type="string" indexed="true" stored="true" required="true" default="view" />

```

The combination of [type](#) and id determines which resource (either community, collection, item page or file download) has been requested.

Unique stored fields for bitstream downloads

```

<field name="bundleName" type="string" indexed="true" stored="true" required="false" multiValued="true" />

```

Unique stored fields for search queries

```

<field name="query" type="string" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="scopeType" type="integer" indexed="true" stored="true" required="false" />
<field name="scopeId" type="integer" indexed="true" stored="true" required="false" />
<field name="rpp" type="integer" indexed="true" stored="true" required="false" />
<field name="sortBy" type="string" indexed="true" stored="true" required="false" />
<field name="sortOrder" type="string" indexed="true" stored="true" required="false" />
<field name="page" type="integer" indexed="true" stored="true" required="false" />

```

Unique stored fields for workflow events

```

<field name="workflowStep" type="string" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="previousWorkflowStep" type="string" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="owner" type="string" indexed="true" stored="true" required="false" multiValued="true"/>
<field name="submitter" type="integer" indexed="true" stored="true" required="false" />
<field name="actor" type="integer" indexed="true" stored="true" required="false" />
<field name="workflowItemId" type="integer" indexed="true" stored="true" required="false" />

```

Disabling Tracking of Statistics

By default, Statistics are captured by the REST API for all visits (page hits) and downloads that occur via both the User Interface and the REST API.

Disabling statistical tracking currently must be done by modifying the backend's Spring configuration in [dspace]/config/spring/rest/event-service-listeners.xml. In that file, you must **comment out** the "SolrLoggerUsageEventListener" in order to disable all tracking

event-service-listeners.xml

```
<beans>
  ...
  <!-- Comment out this bean, as shown below, to disable all tracking of usage statistics in Solr -->
  <!-- Inject the SolrLoggerUsageEventListener into the EventService -->
  <!--
  <bean class="org.dspace.statistics.SolrLoggerUsageEventListener">
    <property name="eventService" ref="org.dspace.services.EventService"/>
  </bean>
  -->
</beans>
```

After commenting out that bean, you will need to restart Tomcat.

NOTE: This only disables **tracking** statistics in Solr. The "Statistics" link will still appear in the header menu of the User Interface. However, you can limit its visibility by setting it to only be visible to administrative users. Update this configuration in your local.cfg or user-statistics.cfg:

```
# Limit access to Admins, Community/Collection Admins
usage-statistics.authorization.admin.usage = true
```

Statistics Menu link can only be removed by customizing DSpace

At this time, there is no flag to remove the "Statistics" menu link completely. See <https://github.com/DSpace/DSpace/issues/9698>

Web User Interface Elements

Pageview and Download statistics

In the UI, pageview and download statistics can be accessed from the "Statistics" navigation menu near the header. That statistics page is "context aware", so it will show the usage statistics for whatever page (site, Community, Collection) you are currently on.

If you are not seeing the menu, it's likely that they are only enabled for administrators in your installation. Change the configuration parameter "authorization.admin.usage" in usage-statistics.cfg to false in order to make statistics visible for all repository visitors.

Home page

Starting from the repository homepage, the statistics page displays the top 10 most popular items of the entire repository.

Community home page

The following statistics are available for the community home pages:

- Total visits of the current community home page
- Visits of the community home page over a timespan of the last 7 months
- Top 10 country from where the visits originate
- Top 10 cities from where the visits originate

Collection home page

The following statistics are available for the collection home pages:

- Total visits of the current collection home page
- Visits of the collection home over a timespan of the last 7 months
- Top 10 country from where the visits originate
- Top 10 cities from where the visits originate

Item home page

The following statistics are available for the item home pages:

- Total visits of the item
- Total visits for the bitstreams attached to the item
- Visits of the item over a timespan of the last 7 months
- Top 10 country views from where the visits originate
- Top 10 cities from where the visits originate

Search Query Statistics

Only supported in DSpace 6 and below

Search query statistics are only supported in DSpace 6.x and below at this time. The below screenshots and instructions are for 6.x and will need updating if this feature is ported to later versions of DSpace. See <https://github.com/DSpace/DSpace/issues/2880>

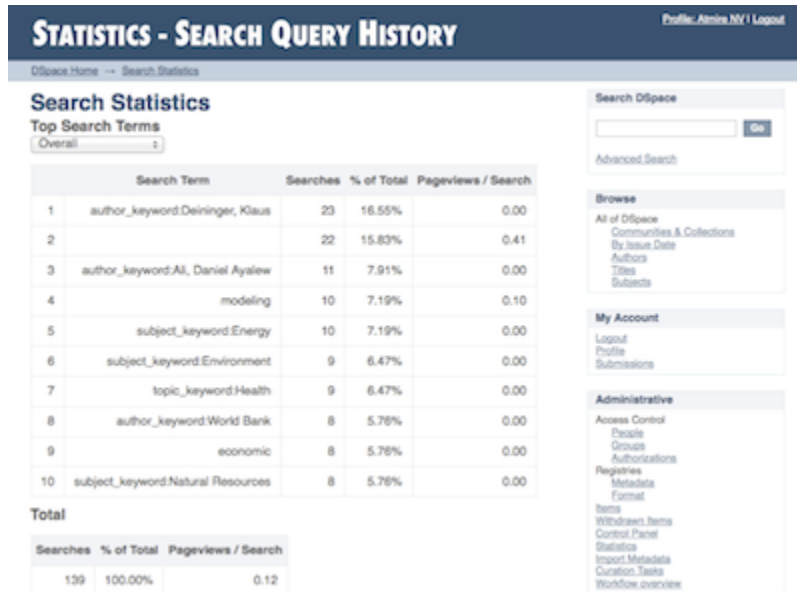
In the UI, search query statistics can be accessed from the lower end of the navigation menu.

If you are not seeing the link labelled "search statistics", it is likely that they are only enabled for administrators in your installation. Change the configuration parameter "authorization.admin.search" in usage-statistics.cfg to false in order to make statistics visible for all repository visitors.

The dropdown on top of the page allows you to modify the time frame for the displayed statistics.

The Pageviews/Search column tracks the amount of pages visited after a particular search term. Therefor a zero in this column means that after executing a search for a specific keyword, not a single user has clicked a single result in the list.

If you are using Discovery, note that clicking the [facets](#) also counts as a search, because clicking a [facet](#) sends a search query to the Discovery index.



Workflow Event Statistics

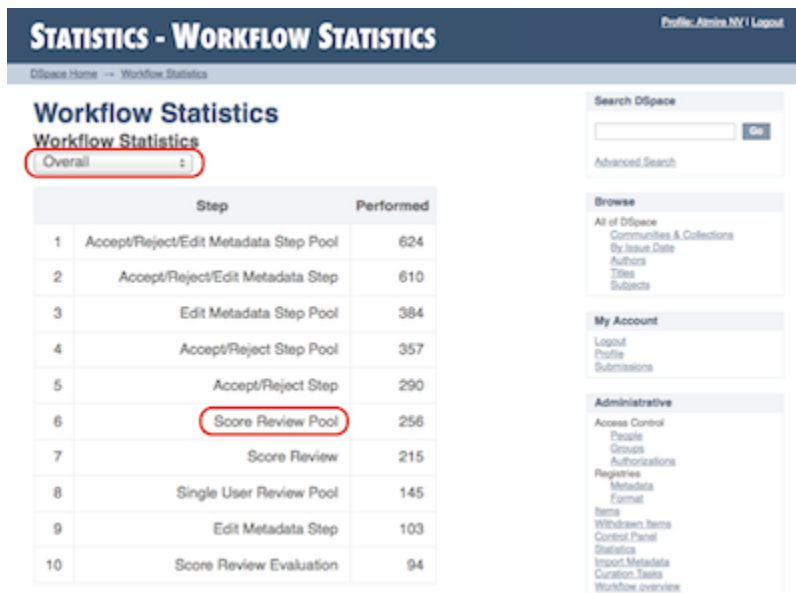
Only supported in DSpace 6 and below

Workflow Event statistics are only supported in DSpace 6.x and below at this time. The below screenshots and instructions are for 6.x and will need updating if this feature is ported to later versions of DSpace. See <https://github.com/DSpace/DSpace/issues/2851>

In the UI, search query statistics can be accessed from the lower end of the navigation menu.

If you are not seeing the link labelled "Workflow statistics", it is likely that they are only enabled for administrators in your installation. Change the configuration parameter "authorization.admin.workflow" in usage-statistics.cfg to false in order to make statistics visible for all repository visitors.

The dropdown on top of the page allows you to modify the time frame for the displayed statistics.



Architecture

The DSpace Statistics Implementation is a Client/Server architecture based on Solr for collecting usage events in the User Interface or REST API applications of DSpace. Solr must be installed separately from DSpace.

Configuration settings for Statistics

In the {dspace.dir}/config/modules/solr-statistics.cfg file review the following fields. These fields can be edited in place, or overridden in your own local.cfg config file (see Configuration Reference).

| | |
|---------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Property: | solr-statistics.server |
| Example Values: | solr-statistics.server = http://127.0.0.1/solr/statistics solr-statistics.server = \${solr.server}/statistics |
| Informational Note: | <p>Is used by the SolrLogger Client class to connect to the Solr server over http and perform updates and queries. In most cases, this can (and should) be set to localhost (or 127.0.0.1).</p> <p>To determine the correct path, you can use a tool like wget to see where Solr is responding on your server. For example, you'd want to send a query to Solr like the following:</p> <div>wget http://127.0.0.1/solr/statistics/select?q=**</div> <p>Assuming you get an HTTP 200 OK response, then you should set solr.log.server to the '/statistics' URL of 'http://127.0.0.1/solr/statistics' (essentially removing the "/select?q=" query off the end of the responding URL.)</p> |
| Property: | solr-statistics.query.filter.bundles |
| Example Value: | solr-statistics.query.filter.bundles=ORIGINAL |
| Informational Note: | A comma seperated list that contains the bundles for which the file statistics will be displayed. |
| Property: | solr-statistics.query.filter.spiderlp |
| Example Value: | solr-statistics.query.filter.spiderlp = false |

| | |
|---------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Informational Note: | If true, statistics queries will filter out spider IPs -- use with caution, as this often results in extremely long query strings. |
| Property: | <code>solr-statistics.query.filter.isBot</code> |
| Example Value: | <code>solr-statistics.query.filter.isBot = true</code> |
| Informational Note: | If true, statistics queries will filter out events flagged with the "isBot" field. This is the recommended method of filtering spiders from statistics. |
| Property: | <code>solr-statistics.autoCommit</code> |
| Example Value: | <code>solr-statistics.autoCommit = true</code> |
| Informational Note: | If true (default), then all view statistics will be committed to Solr whenever the next autoCommit is triggered. This is recommended behavior. If false, then view statistics will be committed to Solr <i>immediately</i> (i.e. via an explicit commit call). This setting is untested in Production scenarios, and is primarily used by automated integration tests (to verify that the statistics engine is working properly). |
| Property: | <code>solr-statistics.spiderips.urls</code> |
| Example Value: | <code>solr-statistics.spiderips.urls =</code> <pre> http://iplists.com/google.txt, \ http://iplists.com/inktomi.txt, \ http://iplists.com/lycos.txt, \ http://iplists.com/infoseek.txt, \ http://iplists.com/altavista.txt, \ http://iplists.com/excite.txt, \ http://iplists.com/misc.txt </pre> |
| Informational Note: | <p>List of URLs to download spiders files into [dspace]/config/spiders. These files contain lists of known spider IPs and are utilized by the SolrLogger to flag usage events with an "isBot" field, or ignore them entirely.</p> <p>The "stats-util" command can be used to force an update of spider files, regenerate "isBot" fields on indexed events, and delete spiders from the index. For usage, run:</p> <pre>dspace stats-util -h</pre> <p>from your [dspace]/bin directory</p> |

In the `{dspace.dir}/config/modules/usage-statistics.cfg` file review the following fields. These fields can be edited in place, or overridden in your own local.cfg config file (see [Configuration Reference](#)).

| | |
|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Property: | <code>usage-statistics.dbfile</code> |
| Example Value: | <code>usage-statistics.dbfile = \${dspace.dir}/config/GeoLite2-City.mmdb</code> |
| Informational Note: | <p>References the location of the installed GeoLite or DB-IP City "mmdb" database file. This file is utilized by the LocationUtils to calculate the location of client requests based on IP address.</p> <p>NOTE: This database file MUST be downloaded, installed and updated using third-party tools. See the "Managing the City Database File" section below.</p> |
| Property: | <code>usage-statistics.resolver.timeout</code> |

| | |
|---------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Example Value: | usage-statistics.resolver.timeout = 200 |
| Informational Note: | Timeout in milliseconds for DNS resolution of origin hosts/IPs. Setting this value too high may result in solr exhausting your connection pool. |
| | |
| Property: | useProxies (Set in dspace.cfg) |
| Example Value: | useProxies = true |
| Informational Note: | Will cause Statistics logging to look for X-Forward URI to detect clients IP that have accessed it through a Proxy service (e.g. the Apache mod_proxy). Allows detection of client IP when accessing DSpace. [Note: This setting is found in the DSpace Logging section of dspace.cfg] |
| | |
| Property: | usage-statistics.authorization.admin.usage |
| Example Value: | usage-statistics.authorization.admin.usage = true |
| Informational Note: | When set to true, only general administrators, collection and community administrators are able to access the pageview and download statistics from the web user interface. As a result, the links to access statistics are hidden for non logged-in admin users. Setting this property to "false" will display the links to access statistics to anyone, making them publicly available. |
| | |
| Property: | usage-statistics.authorization.admin.search |
| Example Value: | usage-statistics.authorization.admin.search = true |
| Informational Note: | When set to true, only system, collection or community administrators are able to access statistics on search queries. |
| | |
| Property: | usage-statistics.authorization.admin.workflow |
| Example Value: | usage-statistics.authorization.admin.workflow = true |
| Informational Note: | When set to true, only system, collection or community administrators are able to access statistics on workflow events. |
| | |
| Property: | usage-statistics.logBots |
| Example Value: | usage-statistics.logBots = true |
| Informational Note: | When this property is set to false, and IP is detected as a spider, the event is not logged. When this property is set to true, the event will be logged with the "isBot" field set to true. (see solr-statistics.query.filter.* for query filter options) |
| | |

| | |
|---------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Property: | usage-statistics.shardedByYear |
| Example Value: | usage-statistics.shardedByYear = false |
| Informational Note: | When set to "true", the DSpace statistics engine will look for additional Solr Shards (per year) when compiling all usage statistics. Therefore, if you are regularly running "stats-utlis -s" (as documented in the " Solr Sharding By Year " section of the "SOLR Statistics Maintenance" page), then you should set this to "true". By default, it is "false", which tells the statistics engine to only compile usage statistics based on what is found in the current Solr core. |

Pre-1.6 Statistics settings

Only supported in DSpace 6 and below



Search query statistics are only supported in DSpace 6.x and below at this time. See <https://github.com/DSpace/DSpace/issues/2852>

Older versions of DSpace featured static reports generated from the log files. They still persist in DSpace today but are completely independent from the SOLR based statistics.

The following configuration parameters applicable to these reports can be found in dspace.cfg.

```
##### Statistical Report Configuration Settings #####

# should the stats be publicly available?  should be set to false if you only
# want administrators to access the stats, or you do not intend to generate
# any
report.public = false

# directory where live reports are stored
report.dir = ${dspace.dir}/reports/
```

These fields are not used by the new 1.6 Statistics, but are only related to the Statistics from previous DSpace releases

Statistics Administration

Converting older DSpace logs into SOLR usage data

If you have upgraded from a previous version of DSpace, converting older log files ensures that you carry over older usage stats from before the upgrade.

Statistics Client Utility

The command line interface (CLI) scripts can be used to clean the usage database from additional spider traffic and other maintenance tasks. As of DSpace 3.0, a script has been added to split up the monolithic SOLR core into individual cores each containing a year of statistics.

Anonymizing Statistics

DSpace provides a commandline script (./dspace anonymize-statistics) which allows you to anonymize your statistics to better comply with GDPR and similar privacy regulations.

The script will anonymise the IP values by rewriting ('masking') the last part. This mask is configurable, both for ipv4 and ipv6 addresses.

- For IPv4 addresses, the last number will be replaced by the mask, defined by the configuration key 'anonymise_statistics.ip_v4_mask' which defaults to '254'.
For example, 109.74.16.171 is rewritten as 109.74.16.254
- For IPv6 address, the last two numbers will be replaced by the mask, defined by the configuration key 'anonymise_statistics.ip_v6_mask' which defaults to 'FFFF:FFFF'. For example, 2001:0db8:85a3:0000:0000:8a2e:0370:7334 is rewritten as 2001:0db8:85a3:0000:0000:8a2e:FFFF:FFFF

For each anonymised record, the DNS field is also replaced by "anonymised".

Script options available:

- The program only processes records older than 90 days. This period can be altered with the config 'anonymise_statistics.time_limit' (expressed in days) in usage-statistics.cfg.
- "-s [sleep]" : The script takes an optional parameter '-s [sleep]' (expressed in ms), which will make the Java thread sleep between the calls to Solr to reduce the load impact.
- "-t [threads]" : The Solr service commit mechanism is also optimised by adding multi-threading support. The script takes an optional parameter '-t [threads]' to indicate how many threads the Solr service can use for this, if not given the thread count defaults to 2.

Statistical records can also be anonymised the moment they are created. Enabling this feature can be done by setting the configuration parameter "anonymise_statistics.anonymise_on_log" to true in "usage-statistics.cfg" When this configuration property is not set, the feature is disabled by default.

Custom Reporting - Querying SOLR Directly

When the web user interface does not offer you the statistics you need, you can greatly expand the reports by querying the SOLR index directly.

Resources

- <https://www.safaribooksonline.com/library/view/apache-solr-enterprise/9781782161363/>
- <https://lucidworks.com/blog/faceted-search-with-solr/>

Examples

Top downloaded items by a specific user

Query:

```
http://localhost:8983/solr/statistics/select?indent=on&version=2.2&start=0&rows=10&fl=%2Cscore&qt=standard&wt=standard&explainOther=&hl.fl=&facet=true&facet.field=epersonid&q=type:0
```

Explained:

facet.field=epersonid — You want to group by epersonid, which is the user id.
type:0 — Interested in bitstreams only

```
<lst name="facet_counts">
  <lst name="facet_fields">
    <lst name="epersonid">
      <int name="66">1167</int>

      <int name="117">251</int>

      <int name="52">42</int>

      <int name="19">36</int>

      <int name="88">20</int>

      <int name="112">18</int>

      <int name="110">9</int>

      <int name="96">0</int>

    </lst>
  </lst>
</lst>
```

Managing the City Database File

If you wish to record the geographic locations of clients in your DSpace statistics records (e.g. the City or Country where they are accessing your DSpace), you **must** install (and regularly update) one of the following IP to City Database Files (in MMDB format). We recommend installing a City-level database, as it provides more granular location information than a Country-level database (which can only provide the country where the access originated).

- Either install a copy of [MaxMind's GeoLite City database](#) (in MMDB format)
 - Installing MaxMind GeoLite2 is *free*. However, you **must** sign up for a (free) MaxMind account in order to obtain a license key to use the GeoLite2 database.
 - You will need to arrange regular downloads of the GeoLite2 database. MaxMind [offers an updater tool \(geoipupdate\)](#) to do the downloading/updating, and a number of Linux distributions package it (as `geoipupdate`). You will still need to configure your license key prior to usage. Use it before restarting DSpace, to get an up-to-date database.
 - Once the "GeoLite2-City.mmdb" database file is installed on your system, you will need to configure its location as the value of `usage-statistics.dbfile` in your `local.cfg` configuration file.
 - NOTE: This file is frequently updated by [MaxMind.com](#), so you will need to refresh it regularly (ideally by scheduling the updater tool via a cron job or similar). As this is written, the database is updated monthly, and to be allowed to obtain it you need to agree to keep your copy updated.
- Or, you can alternatively use/install [DB-IP's City Lite database](#) (in MMDB format)
 - This database is also free to use, but does **not** require an account to download.
 - You will need to arrange regular downloads of the City Lite database. DB-IP [offers an updater tool \(dbip-update\)](#) to do the downloading/updating, but it requires PHP to run.
 - Once the "dbip-city-lite.mmdb" database file is installed on your system, you will need to configure its location as the value of `usage-statistics.dbfile` in your `local.cfg` configuration file.

- NOTE: This file is frequently updated by [DB-IP.com](https://db-ip.com), so you will need to refresh it regularly (ideally by scheduling the updater tool via a cron job or similar). As this is written, the database is updated monthly with the latest available at <https://db-ip.com/db/download/ip-to-city-lite>